CSI Subtask 2.3 Final Report: High-fidelity solar forecasting demonstration for grid integration

Subtask 2.3 - Recommend placement and operation of SDG&E weather stations to maximize forecast effectiveness

Submitted to:

Stephan Barsun, P.E.
Principal Energy Consultant, Itron
California Solar Initiative RD&D Program
and
Brian D'Agostino
San Diego Gas & Electric

March 2014

Submitted by:

Coimbra Energy Group
Department of Mechanical and Aerospace Engineering
Jacobs School of Engineering
Center for Renewable Resource Integration and Center for Energy Research
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0411, USA

Contents:

List of figures:

List of tables:

Description

In this subtask (2.3) we develop a method to estimate solar power variability over the SDG&E territory. The objective is to optimize the deployment of SDG&E's radiometric network in order to facilitate forecasting activities. For a given service territory, defining the number and location of areas with similar/dissimilar solar variability characteristics allows the utility company to strategically site solar generation assets, and to incentivize or discourage solar growth in regions that diminish grid reliability. A resource cluster analysis is designed for planning and operations for future solar growth through the spatial averaging of production fluctuations. Additionally, understanding the variability associated with different regions enables system operators to improve decision-making on unit dispatch by increasing the confidence of unit commitment operations, predicting intra-hour dispatch and reducing automatic generation control (AGC) errors. To this end, we determine the optimal number and spatial distribution of regions of coherent global irradiance based on an unsupervised learning cluster analysis.

1. Data Acquisition

In order to investigate coherent clusters of similar broadband Global Horizontal Irradiance (GHI) while maintaining a uniform spatial discretization over utility-scale areas of interest, irradiance data derived from satellite images is employed for the cluster analysis in this subtask. In particular, we use GHI data from the SolarAnywhere [1] Enhanced Resolution dataset for 2009 and 2010. This dataset consists of GHI derived from the semi-empirical SUNY model which extracts global and direct irradiances from the visible channel of geostationary weather satellites [2]. The spatial and temporal resolution of the dataset are 0.01° x 0.01° (~1 km x 1 km) and 30 minutes respectively. The spatial domain of interest covers the landmass of Southern California between 32°-34°N and 116°-119°W. This utility-scale domain includes the San Diego Gas & Electric service area which supplies power to a population of 1.4 million business and residential accounts in a 4,100 square-mile service area spanning 2 counties and 25 communities.

2. Methodology

2.1 Clear Sky Index

To remove variability associated with deterministic diurnal/seasonal solar cycles, we normalize the GHI time series with respect to a clear sky model (CSM) [3]. Rather than employ a multi-parameter CSM, which can require up to eight atmospheric inputs, in this task we employ the bulk-parameter CSM developed by Ineichen and Perez (2002) [4] which requires only the Linke turbidity coefficient as an input.
The normalized GHI, or clear-sky index K_c as it is more commonly known, is defined as:

$$K_C(t) = \frac{G_h(t)}{G_{hc}(t)} \tag{1}$$

where $G_h(t)$ is the modeled GHI at time t, $G_{hc}(t)$ is the modeled clear-sky GHI at time t, and the dimensionless quantity K_c tends to vary between 0 and 1.

2.2 Variability of the Clearness

As a first step towards coherent GHI clustering, daily variability is considered in this project. To this end, it is desirable to compile daily (rather than intra-hourly) parameters at each pixel to be employed in the cluster analysis. In order to accomplish this, K_c frames corresponding to the same day are averaged at each location to yield vectors of average daily clearness. Note that prior to the normalization, integration of daily GHI values over a day is equivalent to the total daily energy per pixel area (Jm^{-2}) and speaks to the pixel's relative potential for energy production. However, after normalization the integral represents the Daily Average Clear-Sky Index ρ for a given pixel, which is a dimensionless parameter that tends to vary between 0 and 1.
The extracted feature ρ for a daily time course of K_c defined on a closed interval $[0,T_y]$ is approximated by

the trapezoidal rule:

$$\rho = \frac{1}{T_\gamma}\int_{t=0}^{T_\gamma} k_t\, d\,t = \frac{1}{\gamma}[k(0) + 2\sum_{i=1}^{\gamma-1} k(i) + k(\gamma)], \quad (2)$$

where T_γ is the total daylight time (sec) of γ frames for each day, which varies with the day/season.

2.2.1 *Clustering feature #1*

A feature ρ (referred to as the *clustering feature #1*) that represents appropriately the daily average clearness over each precise location of the gridded domain of interest was initially extracted from the normalized GHI data.

This definition of ρ possesses several benefits including, being independent of units, eliminating inconsistencies in the length of days, and is void of deterministic fluctuations resulting in a time series which is dimensionless, bounded, stationary, and completely stochastic.

2.2.2 *Clustering feature #2*

Additionally, a second extracted feature (referred to as the *clustering feature #2*) has been added to this project in order to give emphasis on the fluctuations of the daily clearness index ρ. The variability index is described by the consecutive absolute step changes of the daily average clear-sky clearness index over a precise surface area.

Rather than use the temporal vectors of ρ in the cluster analysis, a final post processing is applied which aims to investigate the variability of clearness at each pixel. To this end, we define a new variability measure of the Daily Average Clear-Sky Index as:

$$Variability_{i=1,\ldots,N} = |[\rho_i(2)-\rho_i(1), \rho_i(3)-\rho_i(2), \ldots, \rho_i(M)-\rho_i(M-1)]|_{i=1,\ldots,N}, \quad (3)$$

where M is the number of days in the dataset for the i^{th} pixel.
We expect to have low values of $\Delta\rho$ for consecutive similar cloud-conditions and high values for a sequence of fluctuating clear and cloudy days.

2.2.3 Feature Vectorization

With a preprocessing technique the original dataset is transformed in a way such that each element of a vector represents each of the two *clustering features* for each location in the area of interest. Figure 1 depicts the described feature extraction stage for the *clustering feature #2*, where the N land-cover pixels are used to construct N temporal vectors that represent the course of $\Delta\rho$ at each pixel's location.

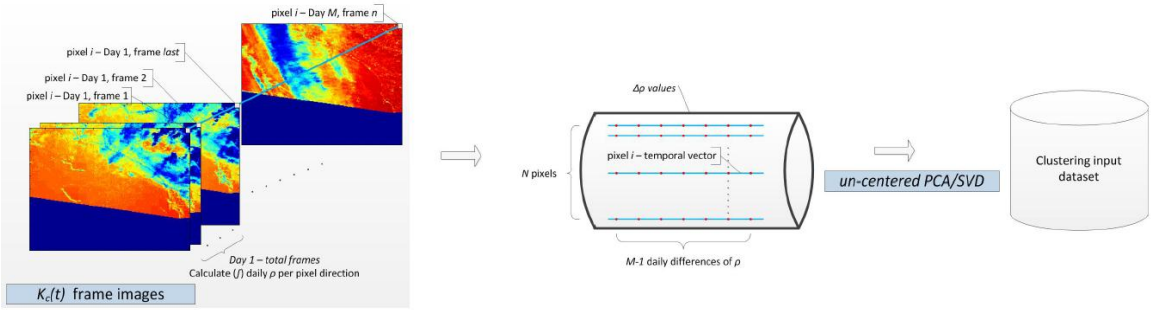

Figure 1 - A sequence of consecutive K_c frames over Southern California. The lower blue segment is not a territory of the United States and is not included in the experiments. An example of how the concatenated pixels over a specific location are used to calculate the daily ρ value is illustrated (*left*). All the temporal vectors (blue lines) that represent the *M-1* daily differences of ρ (red dots) for each of the N pixel locations are then stacked to create the $\Delta\rho$ feature's

dataset (middle). A dimensionality reduction method (un-centered PCA/SVD) precedes to provide the clustering input dataset (*right*).

2.2.4 Dimensionality Reduction

The high dimensionality of the temporal vectors requires a dimensionality reduction in order to lower the computational complexity and remove noise from data. For that purpose we employ the most widely used linear dimensionality reduction method, the Principal Component Analysis (PCA) [5]. In this subtask, the PCA via Singular Value Decomposition (SVD) projects the initial high dimensional un-centered data (no mean-centering) into the best low-dimensional linear approximation in such a manner that 99% of the initial variance of the data is preserved [18].

3. Clustering Approach

The *k-means* clustering algorithm [6] is applied in conjunction with a stable initialization method to diminish its dependency to random initial conditions. Clustering validation is performed by the computation of two internal validity indices [7,8], in order to investigate the number of clusters that best captures the cohesion and separation of the clustering partition with respect to the parameterization of the variability distribution problem. The appropriate number of clusters is estimated by a simple and efficient graphical method, known as the L-method [9].

3.1 Clustering by the *k-means* and Initialization Method

The *k-means* algorithm is the most widely used and simple partitioning clustering method of unsupervised learning [10, 11]. The reasons for its popularity rely primarily on its scalability and simplicity. On the other hand, the algorithm also suffers from a number of limitations. Primarily, it considers the underlying structure of the data as hyper-spherical, owing to the typical selection of the Euclidean distance as the primary clustering criteria. For this reason, *k-means* partitions may be fallacious for dataset structures composed by non hyper-spherical shapes. In addition, the requirement to define a priori the number of *k* clusters can also be considered as a primal handicap. The default iterative refinement algorithm [12] of *k-means* uniformly chooses a random number of *k* points as the initial centers of the desired *k* clusters, where each point of the dataset is assigned to its closest center. Subsequently, the position of the *k* centers is iteratively optimized by the minimization of the distance criteria between the points of a cluster to its center. The algorithm stops either after a predefined number of iterations or when a convergence threshold value of a criterion function is reached. Hence, it is obvious that k-mean's effectiveness depends on how close the initial centers are to the final partition. The initialization of different seed centers generates divergent final clustering solutions. In addition, the risk of convergence to local minima of the criterion distance is high.

In order to achieve a stable solution, several heuristic algorithms have been proposed. Celebi et al. (2013) [13] presented a comprehensive survey along with a comparative study of *k-means* initialization methods. These methods are mainly distributed by their time complexity and their deterministic or non-deterministic heuristic approach to select the initial centers. In this project, we apply a deterministic initialization scheme that provides stable seed centers with respect to a structural parameter [18]. The first step in the two-step method is to select $m=3k$ points based on the density of the reverse nearest neighbors (RNN) [14]. At the second step, *k* initial centers are finally selected which are spaced at least by a predefined threshold distance and count the maximum number of nearest members in descending order.

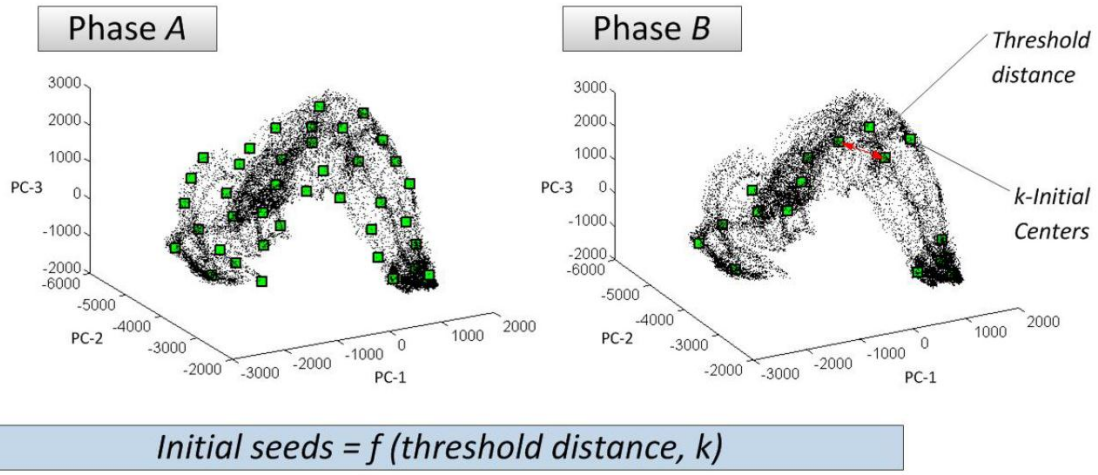

Figure 2 - Three-dimensional plots of the dataset (black dots) used in the experiments in the Principal component (PC) axes. Each of the black dots represents the 3-component feature vector at every spatial location (pixel) of the service area as lying in the direction/subspace that corresponds to the three first PCs. On the left, the first *m = 3k* selected centers at the first step of the initialization method are shown as green square symbols. On the right, a selection of *k* centers that are spaced at least by a predefined distance and count the maximum number of nearest neighbors in descending order are finally used as the cluster centers. Different initial seeds are acquired by applying different threshold distances and *k* number of centers.

3.2 Internal Clustering Validity Indices

The criteria used to estimate how well a proposed clustering fits the structure underlying the partitioned data are called cluster validity indices. In the case that no correct or known partition is available, the clustering validation is achieved by estimating internal measures of the data such as the compactness and the inter separation of the clusters. These types of criteria are known as internal cluster validity indices. Milligan and Cooper (1985) [16] compared 30 validity indices that existed by that time and constitutes an important reference in cluster analysis. Recently, Arbelaitz et al. (2013) [17] published an extensive comparative study of popular and self-subsistent cluster validity indices in different experimental configurations and suggest guidelines to select the most suitable for any particular environment. Typically these indices are based on the computation of the intra-cluster cohesion and the inter-cluster separation. The indices can then estimate the partitions quality in terms of different variations of ratio-type and summation-type factors. These factors are commonly related to the geometrical or statistical properties of the clusters, the similarity or dissimilarity between the data, the number of partitioned data, and/or the number of clusters.

Indeed, a good clustering is equivalent to close quantifiable distances among the member points of a cluster and at the same time, high distinction from the points of other clusters. Most of these methods tend to consider the clusters as hyper-spherical shapes providing the clusters centers as a benchmark for the measurement of compactness and separation. Taking into account the properties of the most frequently cited internal validity measures, the Calinski-Harabasz [7] and the Silhouette [8] indices are employed as the most suitable for this project among others.

3.3 L-method

Estimating the appropriate number of clusters is one of the most ambiguous steps in cluster analysis. Therefore, we seek the knee point of a validity index curve that corresponds to the number of clusters after which no significant change in value of the considered index occurs. In this project we adopt an efficient knee point detection method, called the L-method [9]. The L-method is used for the location of the crucial point on the evaluation graph of any validity index. The main advantage of the L-method is that it does not require the execution of the clustering algorithm itself. It performs a rapid standalone procedure on the curve of an already implemented validity index graph. Unlike the model-based methods,

the L-method detects the boundary between the pair of straight lines that best fit either side of the curve. For example, the ideal shape of the curve forms an '*L*' which implies a sharp change of the considered index's values to a uniform segment. In this case, two lines are fit to the approximately linear left and right parts of the curve indicating the crucial point of discontinuity of the two lines. It should be noted that no significant change of the validity index prevails at the curve segment following the knee point indicating that the clusters are no longer discrete and they do not contribute to an appropriate partition. Regarding less ideal curves where a monotonically smooth decrease occurs, the point of discontinuity of the pairs of lines that best fit the underlying shape of the curve locates the point after which the curve continues more smoothly than at any other point.

The L-method can be implemented by defining an evaluation graph where the values of a validity index are on the y-axis and the number of clusters on the x-axis. By selecting iterative sequences of points left and right of every possible point that can be considered as a knee point, we create all the possible pairs of fitted lines on either side. The first left sequence of points, *L*, must necessarily be comprised by the first 2 points of the curve whereas the right part, *R*, contains the remaining and so forth. This method covers every possible pair of lines. According to the least squares method, a first-degree polynomial *P* approximates the given points of every line segment linearly.
The L-method determines the appropriate pair of lines that best fit the monotonicity of the curve by minimizing the total root-mean-square error $RMSE_{total}$ calculated as:

$$RMSE_{total} = \frac{c-1}{b-1} RMSE_L + \frac{b-c}{b-1} RMSE_R, \quad (8)$$

where *c* corresponds to the vertical projection on x-axis of the point of discontinuity of the left and right lines, $RMSE_L$ and $RMSE_R$ are their root-mean-square error, respectively, and *b* is defined as the maximum number of clusters. The crucial point $\vartheta \in [3, b-2]$ is defined as:

$$\vartheta = \arg\min_c RMSE_{total}, \quad (9)$$

and indicates the appropriate number of clusters. Figure 3 shows an implementation of the L-method over the evaluation graph of the validity index CH values.

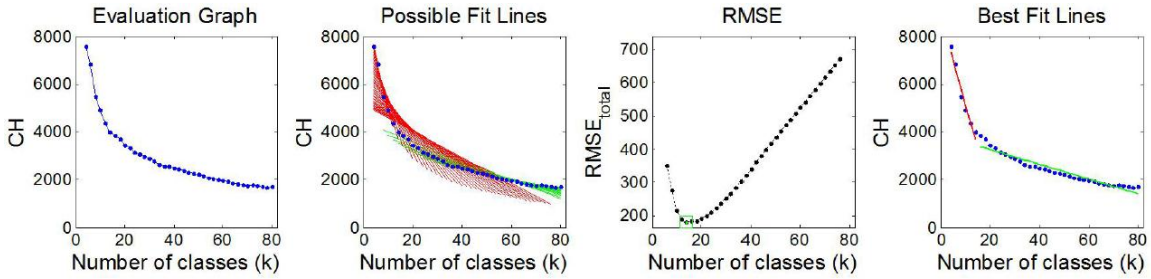

Figure 3 - Determination of the appropriate number of clusters by the implementation of the L-method over the evaluation graph of the CH cluster validity index. From left to right: the diagram of the values of the considered index (y-axis) versus the number of clusters (x-axis), the plot of all possible pairs of fit lines (red and green), the RMSE curve with respect to each candidate knee point (the minimum RMSE, i.e. the knee point, is marked with a green square) and, at the right, the best fit lines where the point of discontinuity defines the knee point.

4. Clustering Maps

The methodology described in the previous section is graphically illustrated in Figure 4. Clustering processes were conducted to a variety of numbers of clusters and the quality of each clustering was evaluated by the validity indices. The appropriate number of clusters is estimated by the L-method where results show convergence to a narrow range of clusters for the two validity indices.

A representative mapping of coherent irradiance clusters is then produced for the region of interest, and the determination of the optimal range of required telemetry sites, and the respective locations for placing solar monitoring installations is calculated.

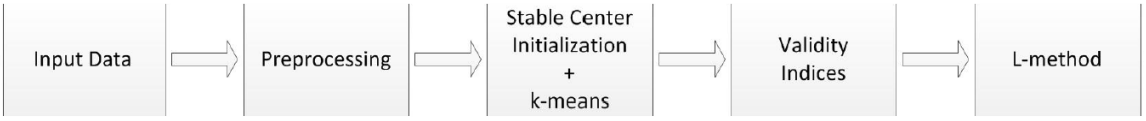

Figure 4 - Block diagram of the proposed methodology

Depending on the *clustering feature* used, a representative spatial segmentation scheme comprises 16 (*clustering feature #*1) or 14 (*clustering feature #*2) coherent clusters in Southern California. Respectively, 8 and 7 of them are located in the SDG&E service area (see Figures 5 and 6). The results show that a narrow range of distinguishable number of clusters derived by the proposed cluster analysis exists, where their centers define the optimum locations of monitoring station installations. Relevant relationships between locations are summarized in Table 1 for 8 cluster centers identified in SDG&E territory and Table 2 for the 7 clusters identified.

Table 1: Results based on *clustering feature #*1, the daily averaged clear sky index (see Fig. 5 for a map). 8 cluster centers are identified. The closest SDG&E met station is listed and (GHI) indicates that that station already has a global horizontal solar irradiance (GHI) sensor. We recommend placement of GHI sensors at location AMO (1), SOB (2), WSY (5), and RAN (8)

	Location of Cluster Center	Closest SDG&E Met Station	Distance and direction to SDGE station. Station is
1	33.3467 , -117.2642	AMO Ammo Dump	0.9 miles North
2	33.0151, -117.2341	SOB RSF	2.5 miles West 2.9 miles East
3	32.814100, -117.0435	MTL Mission Trails (GHI)	2.1 miles North
4	32.995, -116.8829	MGR (GHI)_ CLM	1.6 miles West 1.5 miles North
5	33.0854, -116.6923	WSY	0.9 miles East
6	32.7337, -116.5418	CIR Corte Madera MOR (GHI)	3.4 miles North 3.7 miles South
7	32.8543, -116.1304	IMP (GHI)	9.2 miles
8	33.2161, -116.4615	RAN NRW (GHI)	2.8 miles 11.2 miles

Figure 5 – Irradiance microclimate clusters in the SDG&E service area based on *clustering feature #*1, the daily average clear-sky index. Since the irradiance forecasting skill depends on high quality ground data with high spatial and temporal resolution we recommend installation of additional sensors. Please note that center location 7 has strong topographical constraints.

Table 2: Results based on *clustering feature #*2, the day-to-day variation in averaged clear sky index (see Fig. 6 for a map). 7 cluster centers are identified. The closest SDG&E met station is listed and (GHI) indicates that that station already has a global horizontal solar irradiance (GHI) sensor. All cluster centers have a SDG&E weather station with GHI sensor within 4.1 miles. We recommend palcement of sensors at location TDS (1), MPE (4) and FBK (5).

	Location of Cluster Center	Closest SDG&E Met Station	Distance and direction to SDGE station. Station is
1	32.633 -116.331	TDS BVDC1 (GHI)	1.2 miles 4 miles
2	32.854 -116.652	WDC NDC (GHI)	1.3 miles 2 miles
3	32.884 -116.853	BVY RIO (GHI)	1.3 miles 3.2 miles
4	32.834 -117.174	MPE MSP (GHI)	3.5 miles 4.1 miles
5	33.327 -117.194	FBK CIR (GHI)	2.5 miles 4 miles
6	33.578 -117.445	ORT (GHI)	3.8 miles
7	33.457 -117.615	SCR (GHI)	2.1 miles

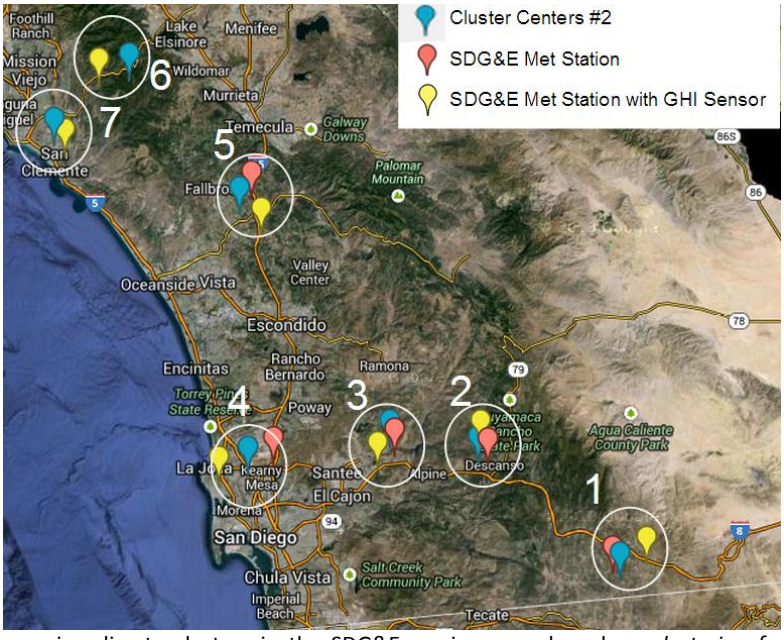

Figure 6 - Irradiance microclimate clusters in the SDG&E service area, based on *clustering feature #*2 (7 cluster centers).

5. Validation with Ground Measurements

Additionally, a validation of the proposed clustering that was applied to satellite solar resource data from Solar Anywhere is performed through a comparison to available ground-based radiometric stations in the

region of interest. Hourly ground measurements from meteorological stations in the California Irrigation Management and Information System (CIMIS) were used. The purpose of this analysis is to demonstrate that data collected by instruments (which are located at ground level within each of the areas of the clusters area) are consistent with the clustering of the satellite- derived time series. A correlation analysis is conducted between the satellite and ground measured GHI time series of 13 CIMIS stations that are located within the examined region of interest in this analysis (see Fig. 7). Ground-based time series from three clusters are selected to be compared: the San Diego region, the Southern Los Angeles Basin, and the Imperial and Coachella Valleys to the east.

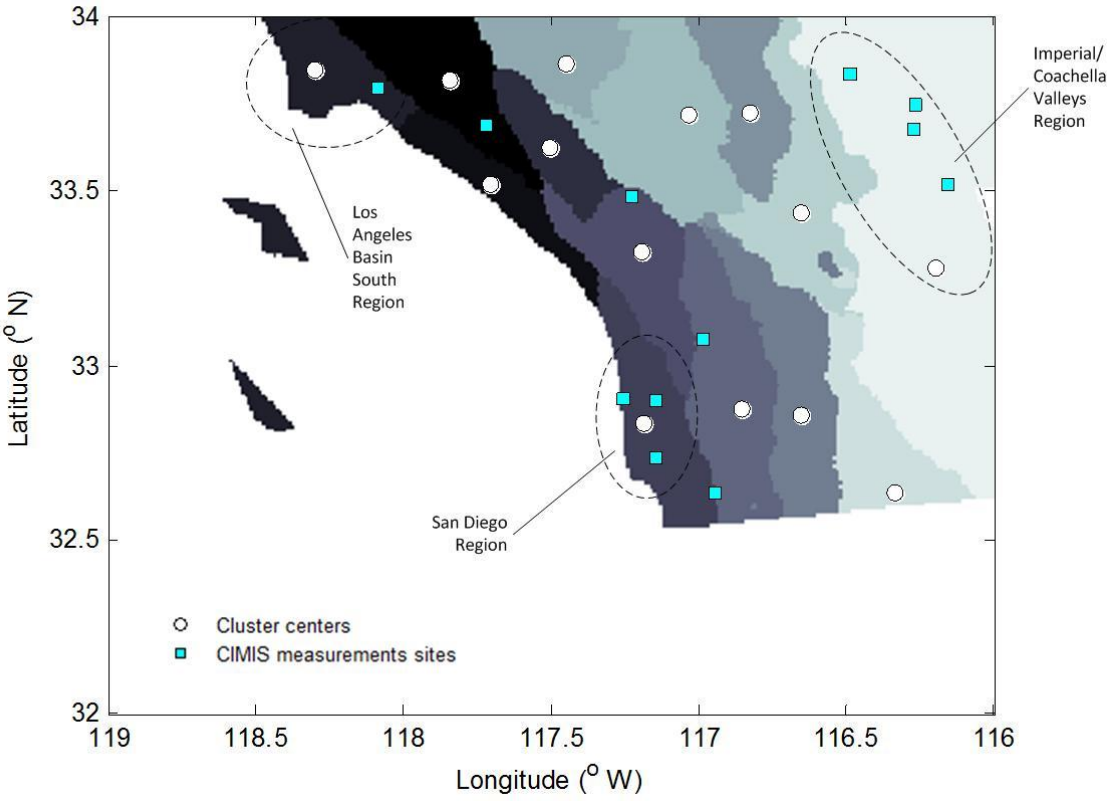

Figure 7 - Segmentation map of the Southern California region into 14 identified irradiance microclimate clusters. Each cluster corresponds to ground locations and is indicated by different color according a decreasing order of the variability expression (i.e. the darker color, the most variant average variability). The proposed centers of each cluster are shown as white circles at their interior and can be considered as ideal sites for potential solar monitoring. The distribution of CIMIS ground-based measurement stations are also indicated (solid blue squares).

The Pearson correlation coefficient results show that high positive correlations (red bands) are clearly observed only between locations within the same cluster region (Fig. 8).

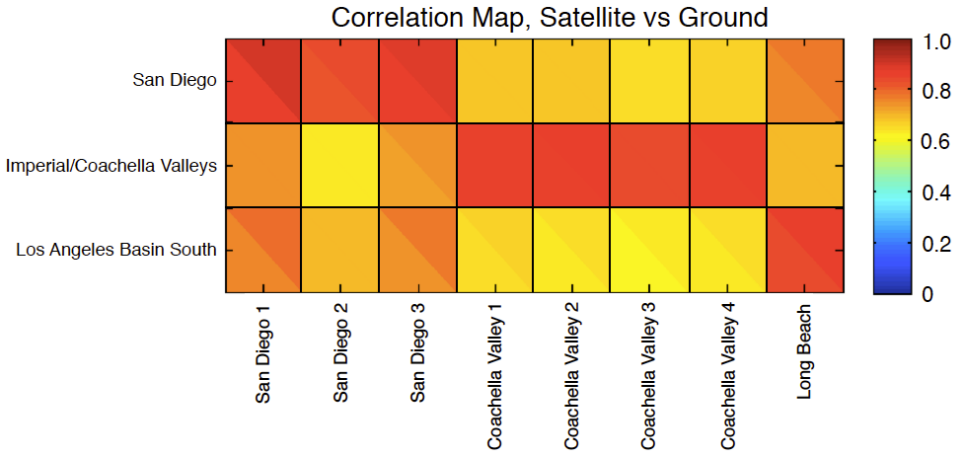

Figure 8 - Cross-correlation between the CIMIS ground stations (columns) and the satellite-derived data (rows) that are depicted in the clustering map (Fig. 7). The top right color of each coordinate corresponds to a correlation between the proposed cluster center and the CIMIS data while the lower left color corresponds to a correlation between 100 randomly selected pixels in the proposed cluster and the CIMIS data. A red value indicates strong correlation (>0.8) between the GHI time series of each of the three cluster centers (see Fig. 7, white circles) and the CIMIS ground stations located within the considered cluster (see Fig. 7, solid blue squares), whereas a yellow value (~0.6) refers to weaker correlations.

Another validation method used for clustering evaluation was to perform a cross-correlation analysis between the time series of the *clustering feature #2* (i.e the *Δρ* time series) to evaluate the intra/inter cluster relationship of the proposed clustering distribution. A high intra-cluster correlation is an indication of excellent coherence within the clusters. Similarly, low inter-cluster correlation argues adequate separation between the clusters. In Figure 9 (*left*), the average correlation coefficients among all the gridded locations belonging to a cluster are shown on the diagonal. The average correlation coefficients between all the possible combinations of inter-cluster correlation are depicted on the off-diagonal segments. Each of the clusters was numbered in order of decreasing variability (see Fig. 9 *right*). It is apparent that the average intra-cluster correlation remains high (>0.9), while the inter-cluster relationship is diminished. Knowledge of the spatial boundaries of correlated/uncorrelated clusters would grant a utility's planning and operations insight into locations of optimal production potential coupled with uncorrelated variability allowing for optimal spatial averaging of resource fluctuations.

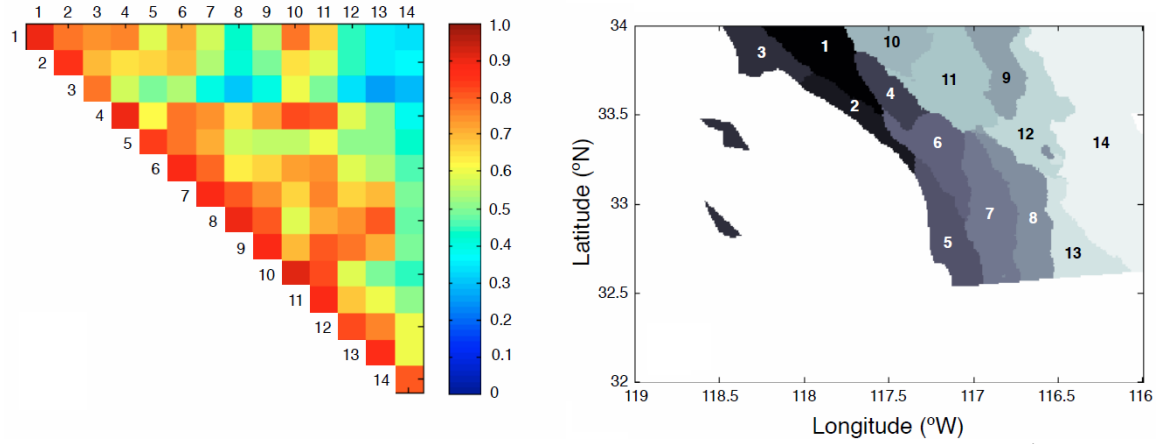

Figure 9 – *Left:* Cross-correlation analysis between the *Δρ* time series in order to evaluate the intra/inter cluster relationship of the proposed clustering. Each column and row of the correlation map is numbered so as to correspond to the cluster labels (*right*), according a decreasing order of the variability expression. A high intra-cluster correlation is an indication of excellent coherence within the clusters just as low inter-correlation argues adequate separation between the clusters. It can be seen that the average intra-cluster correlation remains high (>0.9), while the inter-

cluster relationship is diminished.

6. Summary

The main accomplishments and conclusions of this subtask are as follows:

- Determination of the optimal number and the spatial distribution of regions of coherent global irradiance variability based on an unsupervised learning cluster analysis.
- Depending on the *clustering feature*, there are 16 or 14 different solar variability microclimates in Southern California. The choice between 16 or 14 clusters depends on *clustering feature* #1 or #2, respectively.
- 8 (or 7) of these clusters have their centers located within the SDG&E service area.
    - 1 cluster center is in an area with strong topographical constraints.
- We recommend installation of GHI sensors at locations with no sensor in proximity of 4 miles of the cluster center (AMO , SOB , WSY , RAN , TDS , MPE and FBK).
- All sample rates should be lowered to 30s to resolve temporal variability by clouds.
- A comparison with available ground measurements confirms the coherence of the clustering.
- Knowledge of areas with similar/dissimilar average daily variability can inform improved planning and subsidies for (i) strategic siting of both centralized and distributed solar generations and (ii) effectively reducing variability through coherently optimal spatial averaging.
- The conducted methodology can efficiently be applied to any gridded dataset and any service territory (i.e. both larger and smaller spatial domains as well as finer and coarser temporal scales).

Corresponding Deliverable:

A detailed description of the methods used in this subtask can be found in Ref. [18].

7. References

[1] Solar Anywhere, 2012. SolarAnywhere data, Clean Power Research 2012. http://www.solaranywhere.com

[2] R. Perez, P. Ineichen, K. Moore, M. Kmiecik, C. Chain, R. George, F. Vignola. A new operational model for satellite-derived irradiances: description and validation. Sol. Energy, 73 (5) (2002), pp. 307–317

[3] R.H. Inman, H.T.C. Pedro, C.F.M. Coimbra. Solar forecasting methods for renewable energy integration. Prog. Energy Combust. Sci., 39 (2013), pp. 535–576

[4] P. Ineichen, R. Perez. A new airmass independent formulation for the linke turbidity coefficient. Sol. Energy, 73 (3) (2002), pp. 151–157

[5] I.T. Jolliffe. Principal Component Analysis, vol. 487Springer-Verlag, New York (1986)

[6] J. MacQueen. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 14. California, USA (1967), pp. 281–297

[7] T. Caliński, J. Harabasz. A dendrite method for cluster analysis. Commun. Statist. Theory Methods, 3 (1) (1974), pp. 1–27

[8] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 20 (1987), pp. 53–65

[9] S. Salvador, P. Chan. Learning states and rules for detecting anomalies in time series. Appl. Intell., 23 (3) (2005), pp. 241–255

[10] A.K. Jain. Data clustering: 50 years beyond k-means. Pattern Recognit. Lett., 31 (8) (2010), pp. 651–666

[11] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, D. Steinberg. Top 10 algorithms in data mining. Knowl. Inform. Syst., 14 (1) (2008), pp. 1–37

[12] S. Lloyd. Least squares quantization in PCM. IEEE Trans. Inform. Theory, 28 (2) (1982), pp. 129–137

[13] M.E. Celebi, H.A. Kingravi, P.A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst. Appl., 40 (1) (2013), pp. 200–210

[14] J. Xu, B. Xu, W. Zhang, W. Zhang, J. Hou. Stable initialization scheme for K-means clustering. Wuhan Univ. J. Nat. Sci., 14 (1) (2009), pp. 24–28

[15] S. Theodoridis, K. Koutroumbas. Pattern Recognition (fourth ed.) Academic Press (2009), p. 638

[16] G.W. Milligan, M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50 (2) (1985), pp. 159–179

[17] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona. An extensive comparative study of cluster validity indices. Pattern Recognit., 46 (1) (2013), pp. 243–256

[18] A. Zagouras, R.H. Inman and C.F.M Coimbra (2014), "On the Determination of Coherent Solar Microclimates for Utility Planning and Operations," Solar Energy (102), pp. 173-188